## Title

Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data.



## Journal

PLoS genetics, 5(10)

## ISSN

1553-7390

## Authors

Gutenkunst, Ryan N
Hernandez, Ryan D
Williamson, Scott H
et al.

## Publication Date

2009-10-01

## DOI

10.1371/journal.pgen.1000695

Peer reviewed

OPEN ACCESS Freely available online

PLoS GENETICS

# Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data

**Ryan N. Gutenkunst[1]*, Ryan D. Hernandez[2], Scott H. Williamson[3], Carlos D. Bustamante[3]**

**1** Theoretical Biology and Biophysics and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, **2** Human Genetics, University of Chicago, Chicago, Illinois, United States of America, **3** Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America

## Abstract

Demographic models built from genetic data play important roles in illuminating prehistorical events and serving as null models in genome scans for selection. We introduce an inference method based on the joint frequency spectrum of genetic variants within and between populations. For candidate models we numerically compute the expected spectrum using a diffusion approximation to the one-locus, two-allele Wright-Fisher process, involving up to three simultaneous populations. Our approach is a composite likelihood scheme, since linkage between neutral loci alters the variance but not the expectation of the frequency spectrum. We thus use bootstraps incorporating linkage to estimate uncertainties for parameters and significance values for hypothesis tests. Our method can also incorporate selection on single sites, predicting the joint distribution of selected alleles among populations experiencing a bevy of evolutionary forces, including expansions, contractions, migrations, and admixture. We model human expansion out of Africa and the settlement of the New World, using 5 Mb of noncoding DNA resequenced in 68 individuals from 4 populations (YRI, CHB, CEU, and MXL) by the Environmental Genome Project. We infer divergence between West African and Eurasian populations 140 thousand years ago (95% confidence interval: 40–270 kya). This is earlier than other genetic studies, in part because we incorporate migration. We estimate the European (CEU) and East Asian (CHB) divergence time to be 23 kya (95% c.i.: 17–43 kya), long after archeological evidence places modern humans in Europe. Finally, we estimate divergence between East Asians (CHB) and Mexican-Americans (MXL) of 22 kya (95% c.i.: 16.3–26.9 kya), and our analysis yields no evidence for subsequent migration. Furthermore, combining our demographic model with a previously estimated distribution of selective effects among newly arising amino acid mutations accurately predicts the frequency spectrum of nonsynonymous variants across three continental populations (YRI, CHB, CEU).

**Citation:** Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet 5(10): e1000695. doi:10.1371/journal.pgen.1000695

**Editor:** Gil McVean, University of Oxford, United Kingdom

**Received** February 12, 2009; **Accepted** September 23, 2009; **Published** October 23, 2009



**Funding:** This research was supported by National Science Foundation grant PHY05-51164, National Institutes of Health grants 1R01GM83606 and 2R01HG003229, and DOE contract DE-AC52-06NA25396. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: ryang@lanl.gov

## Introduction

Demographic models inferred from genetic data play several important roles in population genetics. First, they complement archeological evidence in understanding prehistorical events (such as the number and timing of major continental migrations) which have left no written record [1,2]. Second, they facilitate the search for genetic regions that have been targets of non-neutral forces, such as recent natural selection, by guiding our expectations as to how much sequence and haplotype variation one expects to see in a given genomic region (and, more importantly, the variance around these expectations) [3]. Finally, existing demographic models can guide sampling design for subsequent population or medical genetic studies.

Given their many uses, it is not surprising that many studies have inferred demographic models for populations of humans and other species [4–15].

The process of inferring a demographic model consistent with a particular data set typically involves exploring a large parameter space by simulating the model many times, often using coalescent-theory based Monte Carlo approaches. For computational reasons, many of the demographic inference procedures developed thus far have focused on single population models or models with multiple populations but no subsequent migration after subpopulations split (i.e., [4–6,16,17], but also see [10,18]). Methods that do consider multiple populations with migration often assume independent non-recombining regions [7,19] and do not often scale to genomic size data sets. Approaches for jointly considering recombination and migration often use a restricted set of summary statistics [9] of the data, which limits their statistical power. Finally, complex demographic inferences that make use of many summary statistics are often very computationally intensive [8,10,18], which precludes thorough investigation of their statistical properties.

Here, we develop and apply a computationally efficient diffusion-based approach to the problem of demographic inference, based on the multi-population allele frequency spectrum (AFS) (i.e., the joint distribution of allele frequencies across diallelic variants) [10,17,18,20,21]. Given a genetic region sequenced in multiple individuals from each of $P$ populations, the resulting AFS is a $P$-dimensional matrix. Each entry of this matrix records the

## Author Summary

The demographic history of our species is reflected in patterns of genetic variation within and among populations. We developed an efficient method for calculating the expected distribution of genetic variation, given a demographic model including such events as population size changes, population splits and joins, and migration. We applied our approach to publicly available human sequencing data, searching for models that best reproduce the observed patterns. Our joint analysis of data from African, European, and Asian populations yielded new dates for when these populations diverged. In particular, we found that African and Eurasian populations diverged around 100,000 years ago. This is earlier than other genetic studies suggest, because our model includes the effects of migration, which we found to be important for reproducing observed patterns of variation in the data. We also analyzed data from European, Asian, and Mexican populations to model the peopling of the Americas. Here, we find no evidence for recurrent migration after East Asian and Native American populations diverged. Our methods are not limited to studying humans, and we hope that future sequencing projects will offer more insights into the history of both our own species and others.


number of diallelic genetic polymorphisms in which the derived allele was found in the corresponding number of samples from each population. For example, if diploid individuals from two populations were sequenced, with 10 individuals from population 1 and 5 from population 2, the AFS would be a 21-by-11 matrix (indexed from 0). The [2,0] entry would record the number of polymorphisms for which the derived allele was seen twice in population 1 but never seen in population 2, while the [20,5] entry would record polymorphisms for which the derived allele was homozygous in all individuals from population 1 and seen 5 times in population 2. If all polymorphic sites possess only two alleles and can be considered independent, the AFS is a complete summary of the data. Many of the statistics commonly used for population genetic inference, such as $F_{ST}$ and Tajima's $D$, are summaries of the AFS (see [18,22]).

Efficient techniques exist for simulating the AFS of a single population [4,5,23]. The joint AFS between two populations has been used by several recent studies [10,11,18,24], but these have all relied upon very computationally intensive coalescent simulations. Here we approximate the joint multi-population AFS by numerical solution of a diffusion equation, and our implementation supports up to three simultaneous populations. Because the diffusion approach neglects linkage, our comparison with the data is through a composite likelihood function. Such likelihoods are consistent estimators under a wide range of population genetic scenarios for selectively-neutral data, but do not correctly capture variances [25]. (Lower recombination induces higher linkage and higher variance in the entries of the AFS.) As we demonstrate below, the efficiency of our diffusion approach enables both conventional and parametric bootstrap resampling of the data, allowing us to accurately estimate confidence intervals for parameter values and critical values for hypothesis tests [26], accounting for any degree of linkage found in the data. This bootstrap procedure overcomes the traditional concerns with composite likelihood as a philosophy for inference in population genetics.

To demonstrate the utility of our approach, we apply our method to two epochs in human history, using single nucleotide polymorphism (SNP) data from the Environmental Genome Project (EGP) [27], the largest public database of human resequencing data. We first study the expansion of humans out of Africa, jointly modeling the history of African, European, and East Asian populations. We then study the settlement of the New World, jointly modeling European, East Asian, and admixed Mexican populations. In both cases, we quantify the uncertainty of our parameter inferences and test hypotheses about migration (bootstrapping to account for linkage). In particular, we infer an earlier divergence between African and Eurasian populations than previous studies, because our inferences account for the substantial migration between these populations. Our methods also find no evidence for multiple migrations between East Asia and the New World. While similarly complex models for human continental populations have been studied [8], to our knowledge, our analysis is the first in which the full joint AFS is used for inference and in which uncertainty and goodness-of-fit have been quantified.

An important advantage of the diffusion approach is the ease with which selection can be incorporated. As an illustrative application, we also predict the distribution of protein-coding variation between populations. In agreement with the data, we find that less nonsynonymous variation is shared between populations than might be expected based only on patterns of shared noncoding variation.

While no model can capture the full complexity of any species' genetic history, the models presented refine our understanding of the expansion of humanity across the globe. None of the methodology is specific to humans, and we expect our method will find wide application to demographic inference of other species.

## Methods

### Diffusion approximation

To efficiently simulate the AFS, we adopt a diffusion approach. Such approaches have a long and distinguished history in population genetics, dating back to R. A. Fischer [28–30]. The diffusion approach is a continuous approximation to the population genetics of a discrete number of individuals evolving in discrete generations. An important underlying assumption is that per-generation changes in allele frequency are small. Consequently, the diffusion approximation applies when the effective population size $N$ is large and migration rates and selection coefficients are of order $1/N$.

If we have samples from $P$ populations, the numbers of sampled sequences from each population are $n_1,n_2,\ldots,n_P$. (For diploids, $n_1$ is typically twice the number of individuals sampled from population 1.) Entry $d_1,d_2,\ldots,d_P$ of the AFS records the number of diallelic polymorphic sites at which the derived allele was found in $d_1$ samples from population 1, $d_2$ from population 2, and so forth. (If ancestral alleles cannot be determined, then the "folded" AFS can be considered, in which entries correspond to the frequency of the minor allele.)

We model the evolution of $\phi(x_1,x_2,\ldots,x_P,t)$, the density of derived mutations at relative frequencies $x_1,x_2,\ldots,x_P$ in populations $1,2,\ldots,P$ at time $t$. (All $x$ run from 0 to 1.) Given an infinitely-many-sites mutational model [31] and Wright-Fisher reproduction in each generation, the dynamics of $\phi$ for an arbitrary finite number of populations $P$ are governed by a linear diffusion equation:

$$\frac{\partial}{\partial\tau}\phi = \frac{1}{2}\sum_{i=1,2,\ldots,P}\frac{\partial^2}{\partial^2 x_i}\frac{x_i(1-x_i)}{\nu_i}\phi - \sum_{i=1,2,\ldots,P}\frac{\partial}{\partial x_i}\left(\gamma_i x_i(1-x_i)+\sum_{j=1,2,\ldots,P}M_{i\leftarrow j}(x_j-x_i)\right)\phi. \quad (1)$$

The first term models genetic drift, and the second term models selection and migration. Figure 1A illustrates the effects of different evolutionary forces on components of $\phi$. Time is in units of $\tau = t/(2N_{ref})$, where $t$ is the time in generations and $N_{ref}$ is a reference effective population size. The relative effective size of population $i$ is $\nu_i = N_i/N_{ref}$. The scaled migration rate is $M_{i \leftarrow j} = 2N_{ref} m_{i \leftarrow j}$, where $m_{i \leftarrow j}$ is the proportion of chromosomes per generation in population $i$ that are new migrants from population $j$. (Thus migration is assumed to be conservative [32]). Finally, the scaled selection coefficient is $\gamma_i = 2N_{ref} s_i$, where $s_i$ is the relative selective advantage or disadvantage of variants in population $i$. Boundary conditions are no-flux except at two corners of the domain, where all population frequencies are 0 or 1; these are absorbing points corresponding to allele loss or fixation. Because the diffusion equation is linear, we can solve simultaneously for the evolution of all polymorphism by continually injecting $\phi$ density at low frequency in each population (at a rate proportional to the total mutation flux $\theta$), corresponding to novel mutations.

Changes in population size and migration alter the parameters in Equation 1, while population splits and mergers alter the dimensionality of $\phi$. For example, if new population 3 is admixed with a proportion $f$ from population 1 and $1-f$ from population 2 then

$$\phi(x_1, x_2, x_3, t) = \phi(x_1, x_2, t)\, \delta(x_3 - [f x_1 + (1-f) x_2]), \qquad (2)$$

where $\delta$ denotes the Dirac delta function. To remove population 2, $\phi$ is integrated over $x_2$: $\phi(x_1, x_3, t) = \int_0^1 \phi(x_1, x_2, x_3, t)\, \mathrm{d}x_2$.

Given $\phi$, the expected value of each entry of the AFS, $M[d_1, d_2, \ldots, d_P]$, is found via a $P$-dimensional integral over all possible population allele frequencies of the probability of sampling $d_1, d_2 \ldots, d_P$ derived alleles times the density $\phi$ of sites with those population allele frequencies. For SNP data obtained by resequencing, these probabilities are binomial, so

$$M[d_1, d_2, \ldots, d_P] = \int_0^1 \cdots \int_0^1 \prod_{i=1,2,\ldots,P} \binom{n_i}{d_i} x_i^{d_i} (1-x_i)^{n_i - d_i} \phi(x_1, x_2, \ldots, x_P)\, \mathrm{d}x_i. \qquad (3)$$

In some cases of ascertained data [33], the resulting bias can be corrected by modifying the above equation [11,34].

## Likelihood-based inference

Let $\Theta$ correspond to the parameters of a demographic model we wish to estimate from the observed multi-population allele frequency spectrum, which we denote $S[d_1, d_2, \ldots, d_P]$. Assuming no linkage between polymorphisms, each entry in the AFS is an independent Poisson variable [20], with mean $M[d_1, d_2, \ldots, d_P]$ (which depends on $\Theta$). We can, therefore, construct a likelihood function $\mathcal{L}(\Theta|S)$ using standard statistical theory:

$$\mathcal{L}(\Theta|S) = \prod_{i=1\ldots P} \prod_{d_i = 0 \ldots n_i} \frac{e^{-M[d_1, d_2, \ldots, d_P]} M[d_1, d_2, \ldots, d_P]^{S[d_1, d_2, \ldots, d_P]}}{S[d_1, d_2, \ldots, d_P]!}. \qquad (4)$$

So $\mathcal{L}$ is the product of $(n_1+1)(n_2+2)\ldots(n_P+1)$ Poisson likelihoods, one for each entry in the AFS.

In words, our approach consists of calculating the expected allele frequency spectrum $M$ using a particular demographic model (and set of parameter values for that demographic model) using our diffusion approach. We then maximize the similarity between $M$ and the observed AFS $S$ over the parameter values that $\Theta$ can take on. Competing demographic models can be chosen from using standard statistical theory such as the nested likelihood ratio test or information criteria such as the Akaike or Bayesian Information Criteria.

For linked polymorphisms, $\mathcal{L}$ is a composite likelihood. Such likelihoods are consistent estimators under a wide range of neutral population genetic scenarios [25], but simulations incorporating linkage are necessary to estimate variances and define critical

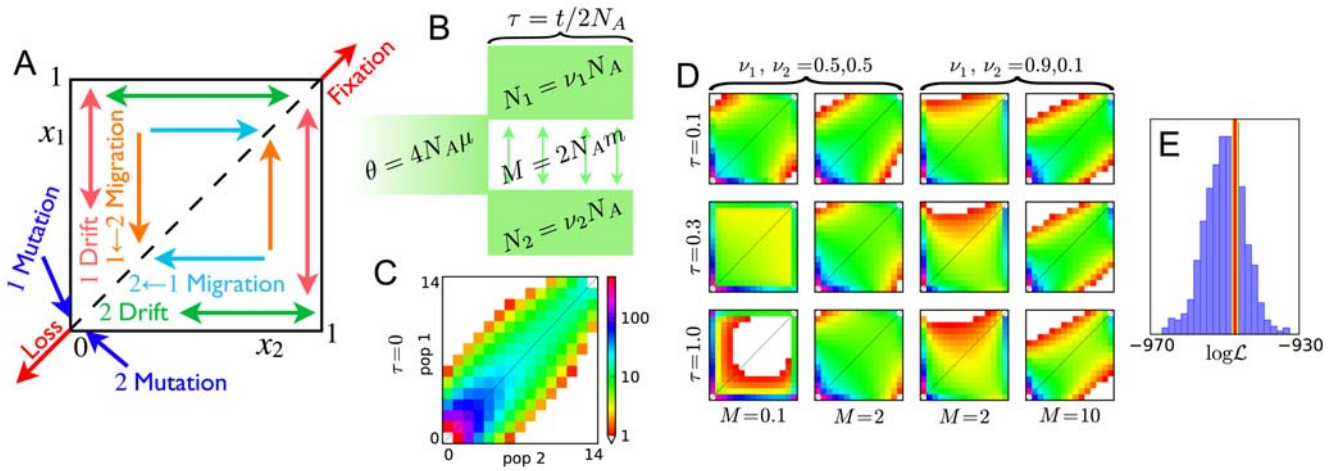


**Figure 1. Frequency spectrum gallery.** (A) Qualitative effects of modeled neutral genetic forces on $\phi(x_1, x_2, t)$, the density of alleles at relative frequencies $x_1$ and $x_2$ in populations 1 and 2. (B) For the spectra shown, an equilibrium population of effective size $N_A$ diverges into two populations $2N_A\tau$ generations ago. Populations 1 and 2 have effective sizes $\nu_1 N_A$ and $\nu_2 N_A$, respectively. Migration is symmetric at $m = M/(2N_A)$ per generation, and $\theta = 1000$. (C) The AFS at $\tau = 0$. Each entry is colored by the logarithm of the number of sites in it, according to the scale shown. (D) The AFS at various times for various demographic parameters, on the same scale as (B). (E) Comparison between coalescent- and diffusion-based estimates of the likelihood $\mathcal{L}$ of data generated under the model (A). Coalescent-based estimates of the likelihood, each of which took approximately 7.0 seconds, are represented in the histogram. The result from our diffusion approach, which took 2.0 seconds, is represented by the red line. For accuracy comparison, the yellow line indicates the likelihood inferred from $10^8$ coalescent simulations.
doi:10.1371/journal.pgen.1000695.g001

values for hypothesis testing and model selection. In our applications, we estimate variances using simulations from the coalescent simulator *ms* [35].

## Numerics

Solving the multi-population diffusion equation is substantially more demanding than the single-population case [23]. This is primarily because the boundary conditions are more complex, and the numerical grid of population frequencies $x$ must be much coarser to be computationally tractable, because it is of $P$ dimensions. For example, a previous single-population study [23] used a uniform $x$ grid of order $10^4$ values between 0 and 1. Extending this grid to a three-population simulation would require an infeasible array of size $10^{12}$. Instead, we use a nonuniform grid and extrapolation to enable accurate computation using of order 100 values along each dimension, for a final array size of order $10^6$.

We solve the diffusion equation on a regular nonuniform grid, using a finite difference scheme [36] inspired by the method of Chang and Cooper [37] (Text S1). Mutations in population $i$ arise at frequency $1/(2N_i) = 1/(2N_{ref}\nu_i)$. The diffusion approximation applies when $N_{ref} \to \infty$, but the minimum frequency in our numerical simulation is that of the first grid point, denoted $\Delta$. To overcome this, we extrapolate our results to an infinitely fine grid. We use a quadratic extrapolation on the logarithm of the AFS entry, modeling the bias introduced by the finite initial grid point $\Delta$ as

$$\log M_{\mathrm{calc}}(\Delta) = \log M_\infty + a\Delta + b\Delta^2. \quad (5)$$

Here $M_{\mathrm{calc}}(\Delta)$ is an AFS element calculated at grid size $\Delta$ and $M_\infty$ is the extrapolated value. Given three evaluations at different grid sizes $\Delta$, we solve for $M_\infty$ and use this value when calculating likelihoods. This vastly increases both the speed and accuracy of our calculation (Supplementary Figure 3 in Text S1). While higher-order extrapolations may improve accuracy in some cases, they may also be more sensitive to numerical noise. Our empirical experience is that a quadratic approximation provides a good compromise between accuracy, efficiency, and robustness.

The computational cost for a single likelihood evaluation scales as $G^{P+1}$ where $G$ is the number of grid points used. In our experience, for stability and accuracy $G$ should somewhat larger than the largest population sample size. Although our theoretical framework extends to an arbitrary number of populations, the exponential scaling of computation with $P$ limits our current applications to three simultaneous populations. Importantly, our likelihood calculation is deterministic and numerically smooth, so numerical derivatives can be used in optimization. We use the the quasi-Newton BFGS method [36], which converges in order $N_\Theta^2$ steps, where $N_\Theta$ is the number of free parameters.

Our implementation of these methods, ∂ a ∂ i, is written in cross-platform Python and C, making use of the NumPy [38], Scipy [39], and Matplotlib libraries [40]. It is distributed under the open-source BSD license. All calculations herein were performed with ∂ a ∂ i version 1.1.0.

We estimated parameter uncertainties by both conventional bootstrap (fitting data sets resampled over loci) and parametric bootstrap (fitting simulated data sets). To generate simulated data we used the coalescent program *ms* [35], a region-specific recombination rate, and the detailed EGP sequencing strategy (Text S1).

The confidence intervals reported in Table 1 and Table 2 derive from a normal approximation to the bootstrap results. For the conventional bootstrap, confidence intervals were calculated as $\overline{\theta^*} \pm 1.96\sigma(\theta^*)$. For the parametric bootstrap, biased-corrected intervals were calculated as $\hat{\theta} - (\overline{\theta^*} - \hat{\theta}) \pm 1.96\sigma(\theta^*)$. The maximum-likelihood value is denoted $\hat{\theta}$, while $\overline{\theta^*}$ and $\sigma(\theta^*)$ denote the mean and standard deviation of the bootstrap results. Aside from the growth rates $r$, all our model parameters are positive by definition, so in those cases we used their logarithms when calculating confidence intervals.

Pearson's $\chi^2$ goodness-of-fit test was performed using all $21^3 - 2 = 9259$ bins in the AFS. Results are similar if we restrict

**Table 1.** Out of Africa inferred parameters.

| parameter[a] | Maximum likelihood | conventional bootstrap 95% confidence interval | parametric bootstrap bias-corrected 95% confidence interval |
|---|---|---|---|
| $N_A$ | 7,300 | 4,400–10,100 | 6,300–9,200 |
| $N_{AF}$ | 12,300 | 11,500–13,900 | 11,100–13,100 |
| $N_B$ | 2,100 | 1,400–2,900 | 1,700–2,600 |
| $N_{EU0}$ | 1,000 | 500–1,900 | 500–1,500 |
| $r_{EU}$ (%) | 0.40 | 0.15–0.66 | 0.26–0.57 |
| $N_{AS0}$ | 510 | 310–910 | 320–750 |
| $r_{AS}$ (%) | 0.55 | 0.23–0.88 | 0.32–0.79 |
| $m_{AF-B}$ ($\times 10^{-5}$) | 25 | 15–34 | 19–36 |
| $m_{AF-EU}$ ($\times 10^{-5}$) | 3.0 | 2.0–6.0 | 1.6–7.6 |
| $m_{AF-AS}$ ($\times 10^{-5}$) | 1.9 | 0.3–10.4 | 0.7–6.9[b] |
| $m_{EU-AS}$ ($\times 10^{-5}$) | 9.6 | 2.3–17.4[b] | 5.7–20.2 |
| $T_{AF}$ (kya) | 220 | 100–510 | 90–410 |
| $T_B$ (kya) | 140 | 40–270 | 60–310 |
| $T_{EU-AS}$ (kya) | 21.2 | 17.2–26.5 | 17.6–23.9 |

[a]See Figure 2B for model schematic. Growth rates $r$ and migration rates $m$ are per generation.
[b]One low-migration outlier was removed for each of these estimations.
doi:10.1371/journal.pgen.1000695.t001

**Table 2.** Settlement of New World inferred parameters.

| parameter[a] | maximum likelihood | conventional bootstrap 95% confidence interval | parametric bootstrap bias-corrected 95% confidence interval |
|---|---|---|---|
| $N_{EU0}$ | 1,500 | 700–2,100 | 900–2,200 |
| $r_{EU}$ (%) | 0.23 | 0.08–0.45 | 0.16–0.34 |
| $N_{AS0}$ | 590 | 320–800 | 410–790 |
| $r_{AS}$ (%) | 0.37 | 0.16–0.60 | 0.24–0.51 |
| $N_{MX0}$ | 800 | 160–1,800 | 140–1,600 |
| $r_{MX}$ (%) | 0.50 | 0.14–1.17 | 0.41–0.98 |
| $m_{EU-AS}$ ($\times 10^{-5}$) | 13.5 | 7.5–32.2 | 9.9–20.8 |
| $T_{EU-AS}$ (kya) | 26.4 | 18.1–43.1 | 21.7–30.7 |
| $T_{Mx}$ (kya) | 21.6 | 16.3–26.9 | 18.6–24.7 |
| $f_{MX}$ (%) | 48 | 42–60 | 41–55 |

[a]See Figure 3B for model schematic. Growth rates $r$ and migration rates $m$ are per generation. $f_{MX}$ is the average European admixture proportion of the Mexican-Americans sampled.

doi:10.1371/journal.pgen.1000695.t002

our analysis to entries in which the expected value is greater than 1 or greater than 5.

## Data

We used the National Institute of Environmental Health Science's Environmental Genome Project SNPs database [41], which results from direct Sanger resequencing of environmental response genes in several populations. We considered all diallelic SNPs in 5.01 Mb of sequence from noncoding regions of 219 autosomal genes (Supplementary Table 8 in Text S1). These data have been the subject of many publications, including [17,23,27,42]. As an assessment of quality, additional high-coverage short-read sequencing has recently been performed across 8 samples in this data set. Over 26,000 sites, the SNP concordance between this next-generation sequencing and the original Sanger sequencing averages 99.5% (D. Nickerson, personal communication). Given the high quality of this data set, we do not incorporate sequencing error into our modeling. We believe such correction will be essential in future applications to less accurate short-read sequencing data, as inference based on the frequency spectrum is sensitive to rare alleles.

To estimate the ancestral allele, we aligned to the panTro2 build of the chimp genome [43]. Like other methods based on the unfolded AFS, our analysis is sensitive to errors in identifying the ancestral allele. We statistically corrected the AFS for ancestral misidentification [17], using a context-dependent substitution model [44]. This procedure has been shown to perform better than aligning to multiple species [17]. To account for missing data and ease qualitative comparisons between populations, we projected all spectra down to 20 samples per population [5] (Text S1).

The human-chimp divergence in the data is 1.13%. We assumed a divergence time of 6 My [45] and a generation time of 25 years. This yielded an estimated neutral mutation rate of $\mu = 2.35 \times 10^{-8}$ per site per generation, which is comparable to direct estimates [46]. There is some controversy as to the appropriate generation time to assume in human population genetic studies [47,48]. In particular, the human generation time may differ between cultures and may have changed during our biological and cultural evolution. The bootstrap uncertainties reported in Table 1 and Table 2 do not include systematic uncertainties in the human-chimp divergence or generation times. The generation time, however, formally cancels when converting between genetic and chronological times.

## Nonsynonymous polymorphism

In our prediction of the distribution of nonsynonymous polymorphism, the distribution of selective effects assumed was a negative-gamma distribution with shape parameter $\alpha = 0.184$ and scale $\beta = 8200$ [49]. The AFS was calculated by trapezoid-rule integration over this distribution, using 201 evaluations logarithmically spaced over $\gamma = [-300, -10^{-6}]$. All demographic parameters, including the scaled mutation rate $\theta$, were set to the maximum-likelihood values from our Out of Africa analysis.

# Results

First, we explored how various demographic forces affect the AFS, building intuition for our subsequent applications to real data. We then compared the performance of diffusion versus coalescent methods for evaluating the AFS, finding that the diffusion approach is substantially faster. We then applied our diffusion approach to infer parameters for plausible demographic models for the history of continental human populations. We first considered the expansion of humans out of Africa and then the settlement of the New World. In these applications, we inferred the maximum composite-likelihood parameters of our models using diffusion fits to the real data. To account for linkage in estimating variances and critical values for hypothesis tests, we then repeatedly fit both conventional and parametric bootstrap data sets. Finally, in an application incorporating selection, we predicted the distribution of nonsynonymous variation between populations in our Out of Africa model, finding good agreement with the available data.

## Demographic effects on the AFS

In Figure 1, we provide examples of the AFS under different demographic scenarios. Figure 1B illustrates the isolation-with-migration model for which the spectra are calculated. The expected spectrum at zero divergence time is shown in Figure 1C. Figure 1D shows the expected spectrum at various divergence times under various demographic scenarios. Qualitatively, correlation between population allele frequencies declines with increasing divergence time, depopulating the diagonal of the

AFS. On the other hand, migration prolongs and sustains correlation. Less obviously, AFS entries corresponding to shared low-frequency alleles distinguish between increased migration and reduced divergence time (Supplementary Figure 1 in Text S1). Additionally, differences in genetic drift between populations with different effective sizes result in asymmetries in the AFS. These qualitative features of the AFS are also evident in human data. Detailed modeling allows us to quantify our inference regarding the type, timing, and strength of demographic events that are consistent with the data.

## Computational performance

The computer program implementing our method is named $\partial$ a $\partial$ i (Diffusion Approximations for Demographic Inference). It is open-source and freely available at http://dadi.googlecode.com.

Figure 1E compares $\partial$ a $\partial$ i with a coalescent approach to evaluating the likelihood of frequency spectrum data. The coalescent simulator *ms* [35] was used to generate a simulated data set from the model in Figure 1B, with parameters $\nu_1 = 0.9$, $\nu_2 = 0.1$, $M = 2$, $\tau = 2$, $\theta = 1000$, scaled total recombination rate $\rho = 1000$, and 20 samples per population. Coalescent-based estimates of the expected AFS were generated by averaging $10^5$ *ms* simulations, each run with $\theta = 1$ and $\rho = 0$. These estimates were scaled to $\theta = 1000$ for comparison with the simulated data set. (This procedure is substantially faster than simulating with larger $\theta$ and $\rho$.) Each estimate took approximately 7.2 seconds of computation. The histogram in Figure 1E shows the resulting distribution of estimated likelihoods of the data. Shown by the red line in Figure 1E is the result from our diffusion approach (with grid sizes $G = \{40,50,60\}$), which took approximately 2.0 seconds of computation. The yellow line is the likelihood from $10^8$ coalescent simulations, illustrating the high accuracy of our diffusion approach. (Note that the coalescent approach we consider here is not necessarily optimal. We are, however, unaware of any such approach that is competitive in computational speed with the diffusion method.)

The computational advantage of the diffusion method is even larger when placed in the context of parameter optimization. Unlike the coalescent approach, there is no simulation variance, so efficient derivative-based optimization methods can be used. As examples, consider our applications to human data, which involve 20 samples per population. On a modern workstation, fitting a single-population three-parameter model took roughly a minute, while fitting a two-population six-parameter model took roughly 10 minutes. The fits of three-population models with roughly a dozen parameters typically took a few hours to converge from a reasonable initial parameter set. This speed allows us to use extensive bootstrapping to estimate variances, overcoming the limitations of composite likelihood.

## Expansion out of Africa

Our analysis of human expansion out of Africa used data from three HapMap populations: 12 Yoruba individuals from Ibadan, Nigeria (YRI); 22 CEPH Utah residents with ancestry from northern and western Europe (CEU); and 12 Han Chinese individuals sampled in Beijing, China (CHB). Because approaches based on the frequency spectrum are sensitive to miscalling of the ancestral state, we statistically corrected for ancestral misidentification using an approach that accounts for a myriad of mutation and context-dependent biases (such as CpG effects) [17]. To ease qualitative comparison among populations and account for missing data, we projected the data down to 20 sampled chromosomes per population [5]. Because this data set is of very high quality (>99% concordance of sequenced SNPs with next-generation sequencing of the same individuals to high coverage; see Methods), we do not explicitly correct for sequencing errors here. We were left with 17,446 segregating diallelic SNPs from effectively 4.04 Mb of sequence. Figure 2A shows the resulting AFS. For ease of visualization, the top row of Figure 2C shows the two-population marginal spectra.

There are many possible three-population demographic models one could consider for these populations. To develop a parsimonious yet realistic model, we first considered the marginal AFS for each population and each pair of populations. Previous analyses found that the YRI spectrum is well-fit by a two-epoch model with ancient population growth [5,17], and we found this as well (Supplementary Figure 6 in Text S1). Previous analyses of the CEU and CHB populations found that both populations went through bottlenecks [5,11] concurrent with divergence [11]. Such models qualitatively fit our marginal CEU-CHB spectrum (Supplementary Figure 7 in Text S1).

Combining these demographic features yields the model illustrated in Figure 2B. The maximum likelihood values for the 14 free parameters are reported in Table 1. Qualitatively, the resulting model reproduces the observed spectra well, as seen in the second and third rows of Figure 2C. (The correlation between adjacent residuals is due in part to our projection of the data down from a larger sample size (Supplementary Figure 8 in Text S1).) Allowing for asymmetric gene flow yielded very little improvement in fit, as did allowing for growth in the Eurasian ancestral population or allowing the CEU and CHB bottleneck and divergence times to differ (data not shown).

Our composite likelihood function assumes that polymorphic sites are independent. Because it thus overestimates the number of effective independent data points, confidence intervals calculated directly from the composite likelihood function will be too small. To control for linkage, we performed both conventional and parametric bootstraps. Because our sequenced genes are typically well separated, they can be treated as independent, and our conventional bootstrap resampled from the 219 sequenced loci. For the parametric bootstrap, simulated data sets that incorporate linkage and the EGP's sequencing strategy were generated with *ms* [35].

Table 1 reports parameter 95% confidence intervals from both the conventional and bias-corrected parametric bootstraps. The parametric bootstraps yield slightly smaller confidence intervals than the conventional bootstrap, suggesting that some variability in the data has not been accounted for by our simulations. This variability may involve small varied selective forces on the sequenced regions or slight relatedness between sampled individuals. The parametric bootstrap results additionally show that our method possesses very little bias in parameter inference (Supplementary Figure 9 in Text S1).

As seen in Table 1, the times for growth in the African ancestral population and divergence of the Eurasian ancestral population ($T_{AF}$ and $T_B$) have particularly wide confidence intervals, likely a consequence of the high inferred migration rate $m_{AF-B}$ between the African and Eurasian ancestral populations. $T_{AF}$ shows high correlation with the ancestral population size $N_A$, while $T_B$ shows no strong linear correlation with any other single parameter (Supplementary Figure 11 in Text S1). We found that 92 out of our 100 conventional bootstrap fits yield $N_{AS0} < N_{EU0}$, supporting the contention that the CHB population suffered a more severe bottleneck than the CEU population [11] (Supplementary Figure 11 in Text S1).

We used several metrics to assess our model's goodness-of-fit, in additional to visual inspection of the residuals seen in Figure 2C. Figure 2D compares the decay of linkage disequilibrium (LD) in the data and in the parametric bootstrap simulations. The

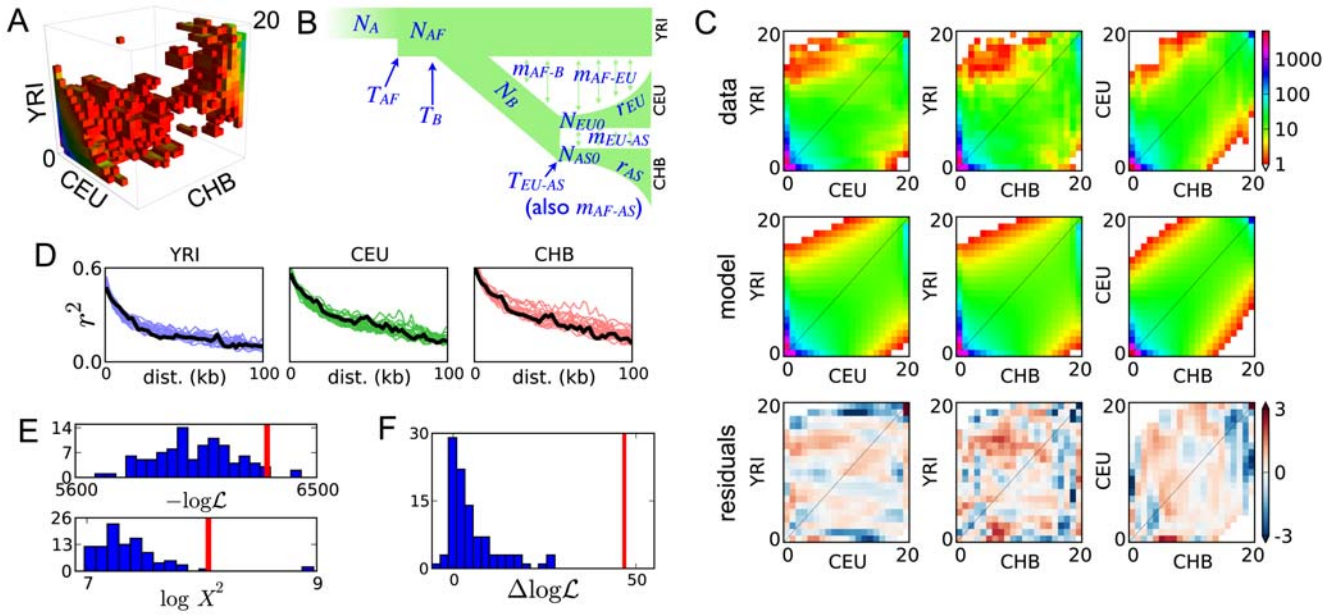


**Figure 2. Out of Africa analysis.** (A) AFS for the YRI, CEU, and CHB populations. The color scale is as in (C). (B) Illustration of the model we fit, with the 14 free parameters labeled. (C) Marginal spectra for each pair of populations. The top row is the data, and the second is the maximum-likelihood model. The third row shows the Anscombe residuals [61] between model and data. Red or blue residuals indicate that the model predicts too many or too few alleles in a given cell, respectively. (D) The observed decay of linkage disequilibrium (black lines) is qualitatively well-matched by our simulated data sets (colored lines). (E) Goodness-of-fit tests based on the likelihood $\mathcal{L}$ and Pearson's $X^2$ statistic both indicate that our model is a reasonable, though incomplete description of the data. In both plots, the red line results from fitting the real data and the histogram from fits to simulated data. Poorer fits lie to the right (lower $\mathcal{L}$ and higher $X^2$). (F) The improvement in likelihood from including contemporary migration in the real data fit (red line) is much greater than expected from fits to simulated data generated without contemporary migration (histogram). This indicates that the data contain a strong signal of contemporary migration.
doi:10.1371/journal.pgen.1000695.g002

agreement seen is notable because our demographic inference used no LD information in building and fitting the model. This LD comparison thus serves as independent validation of both our model and bootstrap simulations. We also asked whether the likelihood $\mathcal{L}$ found in the real data fit is atypical of fits to simulated data. Out of fits to 100 simulated data sets, 2 produced a smaller likelihood (worse fit) than the real data fit (Figure 2E), yielding a p-value of $\approx$0.02. One can craft examples in which a likelihood-based goodness-of-fit test fails to exclude very poor models [50]. Thus we also applied Pearson's $\chi^2$ goodness-of-fit test, a more robust and standard method for data that is in Poisson-distributed bins, such as the AFS [36]. In our case, we must use our parametric bootstraps to assess the significance of the sum-of-squared-residuals test statistic $X^2$, because many entries in the AFS are small and because they are not strictly independent. Figure 2E shows the bootstrap-derived empirical distribution of $X^2$. Two of the bootstraps yielded a larger $X^2$ (worse fit) than the real data fit, giving a p-value of $\approx$0.02, identical to that from the likelihood-based test. (The two simulations that yield a higher $X^2$ than the real fit are not the same two that yield a lower $\mathcal{L}$, suggesting that these tests are somewhat independent.) In some cases specific frequency classes of SNPs, such as rare alleles, may be of particular interest. In Supplementary Table 5 in Text S1, we provide comparisons of the joint distribution of rare alleles seen in the data with that from our simulations. These comparisons indicate that our model also reproduces well this interesting region of the frequency spectrum. Finally, in Figure 4 we compare the model and data using larger bins of SNPs specific to particular populations or segregating at high or low frequency. In all cases the model agrees within the uncertainty of the bootstrapped data. Taken together, these tests suggest that our model provides a reasonable, though not complete, explanation of the data, lending credence to our demographic estimates.

The inferred contemporary migration parameters ($m_{AF-EU}$, $m_{AF-AS}$ and $m_{EU-AS}$) are small, raising the question as to whether they are statistically distinguishable from zero. Figure 2F shows that the improvement in fit to the real data upon adding contemporary migration to the model is much larger than would be expected if there were no such migration, implying that the contemporary migration we infer is highly statistically significant. Omitting ancient migration ($m_{AF-B}$) reduced fit quality even more, indicating that the data also demand substantial ancient migration (data not shown).

## Settling the New World

To study the settlement of the Americas, we used the previously considered 22 CEU and 12 CHB individuals, plus an additional 22 individuals of Mexican descent sampled in Los Angeles (MXL). Data were processed as in our Out of Africa analysis, yielding 13,290 segregating SNPs from effectively 4.22 Mb of sequence. Figure 3A shows the resulting AFS, while Figure 3C shows the marginal spectra.

A model in which the CEU and CHB diverge from an equilibrium population did not reproduce the AFS well (Supplementary Figure 13 in Text S1). Interestingly, a model allowing a prior size change in the ancestral population better fit the AFS but very poorly fit the observed LD decay (Supplementary Figure 13 in Text S1). Thus, reproducing the AFS does not guarantee reproduction of LD, at least given a historically unrealistic model. To develop a more realistic model, we endeavored to include the effects of Eurasian divergence from and migration with the African population. Computational limits precluded us from considering

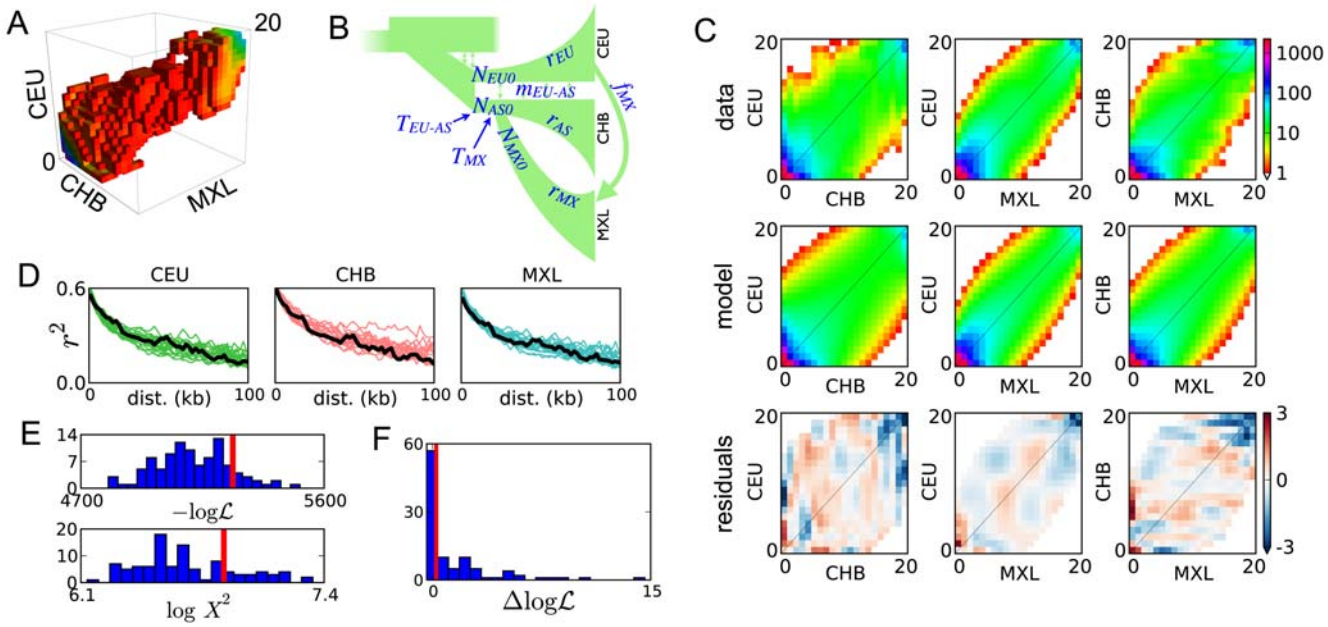


**Figure 3. Settlement of the New World analysis.** As in Figure 2, (A) is the data, (B) is a schematic of the model we fit, (C) compares the data and model AFS, and (D) compares LD. (E) The fit of our model to the real data is not atypical of fits to simulated data. (F) The improvement in real data fit upon including CHB-MXL migration (red line) is very typical of the improvement in fits to simulated data without CHB-MXL migration. Thus we have no evidence for CHB-MXL migration after divergence.
doi:10.1371/journal.pgen.1000695.g003

all 4 populations simultaneously, so we dropped the African population from the simulation upon MXL divergence (Figure 3B).

Table 2 records the maximum-likelihood parameter values inferred for this model. Because this fit did not include African data, we could not reliably infer demographic parameters involving the African population. Thus, for this point estimate we fixed the Africa-related parameters $N_A$, $N_{AF}$, $N_B$, $m_{AF-B}$, $m_{AF-EU}$, $m_{AF-AS}$, $T_{AF}$ and $T_B$ to their maximum-likelihood values from Table 1. Figure 3C compares the model and data spectra. The residuals show little correlation, with the possible exception that the model may underestimate the number of high-frequency segregating alleles.

Parameter confidence intervals are reported in Table 2. To account for our uncertainty in those parameters derived from the Out of Africa fit, for each conventional bootstrap fit we used a set of Africa-related parameters randomly chosen from the sets yielded by our Out of Africa conventional bootstrap. For the parametric bootstrap, we used the maximum-likelihood point estimates. Again, we see that the conventional bootstrap confidence intervals are comparable to, although slightly wider than, the parametric bootstrap intervals. Several parameters in this analysis have direct correspondence with our Out of Africa analysis. Of particular note, the confidence intervals for the CEU-CHB divergence time $T_{EU-AS}$ overlap.

In assessing goodness of fit, Figure 3D shows that this model does indeed reproduce the observed pattern of LD decay. Unlike in our Out of Africa analysis, however, here the LD decay was used to choose the form of the model (although not its parameter values), so this is not a completely independent assessment of fit. Of our 100 parametric bootstrap fits, 13 yielded a worse likelihood than the real fit (Figure 3E), for a p-value of $\approx$0.13. Applying Pearson's $\chi^2$ test, we find that 23 of 100 bootstrap fits yield a higher (worse) $X^2$ than the fit to the real data, for a p-value of $\approx$0.23, similar to that of the likelihood analysis. Comparing distributions of rare alleles, our model typically reproduces the observed distribution well, although it may be somewhat overestimating the proportion of alleles that are rare or absent in the CHB population (Supplementary Table 7 in Text S1). In sum, our model appears to be a reasonable explanation of this data, somewhat better than in our Out of Africa analysis.

An essential feature of the Mexican-American individuals considered here is that they are typically admixed from Native American and European ancestors. The $\approx$50% average European admixture proportion we inferred for the MXL population is consistent with previous estimates for Los Angeles Latinos [51]. We have no direct data from the Native American populations ancestral to MXL, but our model does account for their divergence from East Asia. A model neglecting this divergence (by setting $T_{MX}$ to zero) fit the data substantially worse and yields an unrealistically high average European admixture proportion into MXL of 68%.

Not only are Mexican-American individuals admixed, their admixture proportions also vary, and this subtlety is not directly accounted for in our analysis. To assess its effect on our results, we first roughly estimated the ancestry proportion of each individual, using essentially a maximum-likelihood version [18] of the algorithm used in *structure* [52] (Text S1). (Methods based on "admixture LD", which identify breakpoints between regions of Native American and European ancestry, may be more powerful [53]. However, the strategy used by the EGP of sequencing widely spaced genes will resolve few of these breakpoints, limiting the applicability of these methods.) We then performed additional parametric bootstrap analyses, using simulations with a distribution of individual ancestry chosen to mimic that seen in the data and, to further test the method, with an extremely wide distribution. These simulations showed that variation in individual ancestry does not bias our parameter inferences (Supplementary Figure 19 in Text S1). Remarkably, it does not even change our statistical power. This is evidenced by the fact that these bootstrap simulations yielded confidence intervals identical to our original

simulations without variation in ancestry proportion (Supplementary Figure 19 in Text S1). Nevertheless, future studies may profit by incorporating individual ancestry information [18], perhaps inferred from admixture LD.

Finally, our model allowed us to assess the role recurrent migration from Asia played in the settlement of the New World [2]. When we added CHB-MXL migration to our model, we found that the maximum likelihood migration rate was $1.7 \times 10^{-5}$ per generation. As shown in Figure 3F, the resulting improvement in likelihood is typical (p-value≈0.45) of fits including CHB-MXL migration to data simulated without it. Our data and analysis thus yielded no evidence of recurrent migration in the settlement of the New World. Note, however, that this simple test does not necessarily rule out more complex scenarios, in which migration may vary over time.

### Nonsynomymous polymorphism

Polymorphisms that change protein amino acid sequence are of medical interest because they are particularly likely to affect gene function [54]. Correspondingly, they are often subject to natural selection. Diffusion approaches are particularly useful for studying such nonsynonymous polymorphism, because they easily incorporate selection. Although the diffusion approximation assumes that sites are unlinked, nonsynonymous segregating sites are rare enough that this is often a reasonable approximation [49].

As an illustration, we used our Out of Africa demographic model to predict the distribution of such variation between continental populations. To do so, we must specify a distribution for the selective effects of nonsynonymous mutations that enter the population. For this we adopted a negative gamma distribution whose parameters were recently inferred [49]. The resulting distribution of segregating variation is shown in Figure 4A. (To ease comparison, we have assumed the same scaled mutation rate as in the neutral case of Figure 2C.) As expected, selection sharply reduces the amount of segregating polymorphism. Figure 4B shows the proportion of variants within various classes. Also as expected, selection shifts nonsynonymous variation toward lower frequencies, raising the proportion of singletons and lowering the proportion at frequency greater than 10%. Less obviously, it also reduces the proportion of variation that is shared between populations. In the neutral case, 43% of polymorphism is predicted to be present in more than one population, while in the selected case only 35% is. Thus genetic inferences from coding polymorphism may be less transferable between populations than might be expected from neutral patterns of allele sharing.

In the data considered here, there are about 400 nonsynonymous polymorphisms segregating in the three populations considered. This is too few for a detailed goodness-of-fit test of our predicted distribution. (Although see Supplementary Figure 20 in Text S1 for a direct AFS comparison.) Nevertheless, we observe that our predictions shown in Figure 4B all lie within the bootstrap 95% confidence intervals from the data.

## Discussion

Our diffusion approximation to the joint allele frequency spectrum is a powerful tool for population genetic inference. Although the diffusion approximation neglects linkage between sites, our method's computational efficiency allows us to use extensive bootstrap simulations to account for the effects of linkage. (Let us reiterate that linkage does not affect the expected allele frequency spectrum of neutral sites, so our diffusion-based approach is estimating the same AFS that coalescent simulations are estimating, but in a small fraction of the time). We applied our method to human expansion out of Africa and settlement of the New World, using public resequencing data from the Environment Genome Project. The flexibility of the diffusion approach also allowed us to consider the distribution of non-neutral variation, which is difficult to address with other approaches. Although no model can capture in detail the complete history of any population, the models presented here help refine our understanding of human expansion across the globe.

Our demographic results are in most respects broadly consistent with previous analyses of human populations. In particular, single-population analyses have also inferred African population growth and European and Asian bottlenecks [4–6]. Also, the migration rates we infer are similar to those inferred by Schaffner et al. [8] but somewhat smaller than those of Cox et al. [15]. On the other hand, Keinan et al. [11] inferred no significant migration between CEU and CHB. Finally, our estimate of a New World founding effective population size in the hundreds is compatible other inferences [14].

Perhaps our most interesting demographic results are the inferred divergence times. Other studies [11,12] have estimated divergence times between Europeans and East Asians similar to the ≈23 kya we infer. Interestingly, archeological evidence places humans in Europe much earlier (≈40 kya) [1]. Our inferred divergence time of ≈22 kya between East Asians and Mexican-Americans is somewhat older than the oldest well-accepted New World archeological evidence [2]. The divergence we infer may reflect the settlement of Beringia, rather than the expansion into the New World proper [14]. Finally, the divergence time of ≈140 kya we infer between African and Eurasian populations is consistent with archeological evidence for modern humans in the

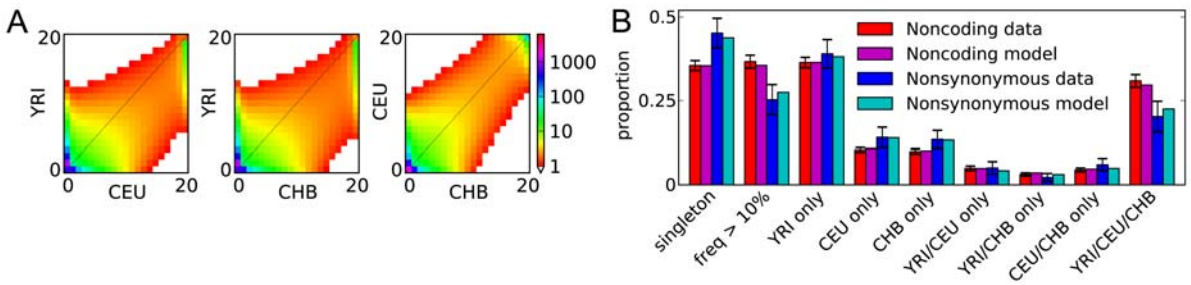


**Figure 4. Distribution of nonsynonymous polymorphism.** We simulated our maximum-likelihood Out of Africa demographic model with a distribution of selective effects previously inferred for nonsynonymous polymorphism [49]. (A) To enable direct comparison with the neutral AFS (Figure 2C), the scaled mutation rate $\theta$ was set identically, as is the color scale. As expected, selection dramatically reduces the amount of segregating polymorphism. (B) Shown are the proportions of variation found in various frequency classes. As expected, nonsynonymous variants typically have lower frequency. They also less likely to be shared between populations. Data error bars indicate 95% bootstrap confidence intervals.
doi:10.1371/journal.pgen.1000695.g004

Middle East ≈100 kya [1], but it is much older than other inferences of ≈50 kya divergence from mitochondrial DNA [1]. This discrepancy may be explained by our inclusion of migration in the model. Migration preserves correlation between population allele frequencies, so an observed correlation across the genome can be explained by either recent divergence without migration or ancient divergence with migration. In fact, the African-Eurasian migration rate we infer of $\approx 25\times10^{-5}$ per generation is comparable to the $\approx 100\times10^{-5}$ inferred from census records between modern continental Europe and Britain [55].

One difficulty in interpreting our divergence times is that the sampled populations may not best represent those in which historically important divergences occurred. For example, the Yoruba are a West African population, so the divergence time we infer between Yoruba and Eurasian ancestral populations may correspond to divergence within Africa itself. Future studies of more populations [56–58] will help alleviate this difficulty.

Another difficulty is that the genic loci we study here may not be ideal for demographic inference. Although we consider only noncoding sequence in fitting our historical model, selection on regulatory or linked coding sites may skew the AFS [59]. In fact, the EGP data have been shown to differ in some ways (e.g. Tajima's *D*) from intergenic regions [58]. Nevertheless, we use the EGP data because it is currently the largest public resource of noncoding human genetic variation, and we fit a neutral model because disentangling the small expected effects of selection on these sites from demographic effects will require additional data. The rapidly declining cost of sequencing will give future studies access to many more loci that are likely to be less influenced by selection. Importantly, the computational burden of our method is independent of the amount of sequence used to construct the AFS. Additional loci will also increase power to discriminate between models and incorporate more detail.

The AFS encodes substantial demographic information. It is has been shown, however, that an isolated population's AFS does not uniquely and unambiguously identify its demographic history [60]; we expect a similar result to hold for multiple interacting populations. Moreover, the AFS does not capture all the information in the data. As illustrated by the alternative New World models we considered, patterns of linkage disequilibrium encode additional information. Future studies may profit from coupling our efficient AFS simulation with methods that address other aspects of the data.

We have developed a powerful diffusion-based method for demographic inference from the joint allele frequency spectrum. We applied our method to human expansion out of African and the settlement of the New World, developing models of human history that refine our knowledge and raise intriguing questions. We also applied our method to predict the distribution of nonsynonymous variation across populations, and this prediction is consistent with the available data. Our methods and the models inferred from it offer a foundation for studying the history and evolution of both our own species and others.

## Supporting Information

**Text S1** Complete supplementary data.
Found at: doi:10.1371/journal.pgen.1000695.s001 (1.80 MB PDF)

## Acknowledgments

We thank Amit Indap for bioinformatics assistance and Jim Booth for statistical assistance. We also thank the NIEHS Program (ES-15478) for making their dataset so easily accessible. We had fruitful discussions with Adam Auton, Debbie Nickerson, Michael Hammer, Rasmus Nielsen, Nick Patterson, Molly Przeworski, Jeff Wall, and Carsten Wiuf.

## Author Contributions

Conceived and designed the experiments: RNG SHW CDB. Performed the experiments: RNG. Analyzed the data: RNG. Contributed reagents/materials/analysis tools: RNG RDH. Wrote the paper: RNG CDB.

## References

1. Mellars P (2006) Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. Science 313: 796–800.
2. Goebel T, Waters MR, O'Rourke DH (2008) The late Pleistocene dispersal of modern humans in the Americas. Science 319: 1497–1502.
3. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857–868.
4. Adams AM, Hudson RR (2004) Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics 168: 1699–1712.
5. Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166: 351–372.
6. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, et al. (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102: 18508–18513.
7. Hey J (2005) On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol 3: e193. doi: 10.1371/journal.pbio.0030193.
8. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–1583.
9. Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. Genome Res 17: 1505–1519.
10. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, et al. (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 3: 1745–1756. doi: 10.1371/journal.pgen.0030163.
11. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet 39: 1251–1255.
12. Garrigan D, Kingan SB, Pilkington MM, Wilder JA, Cox MP, et al. (2007) Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics 177: 2195–2207.
13. Mulligan CJ, Kitchen A, Miyamoto MM (2008) Updated three-stage model for the peopling of the Americas. PLoS ONE 3: e3199. doi: 10.1371/journal.pone.0003199.
14. Kitchen A, Miyamoto MM, Mulligan CJ (2008) A three-stage colonization model for the peopling of the Americas. PLoS ONE 3: e1596. doi: 10.1371/journal.pone.0001596.
15. Cox M, Woerner A, Wall J, Hammer M (2008) Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. BMC Genetics 9: 76.
16. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22: 1185–1192.
17. Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24: 1792–1800.
18. Nielsen R, Hubisz M, Hellmann I, Torgerson D, Andrés A, et al. (2009) Darwinian and demographic forces affecting human protein coding genes.
19. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167: 747–760.
20. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.
21. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159: 1779–1788.
22. Wakeley J (2008) Coalescent Theory: an Introduction Roberts & Company.
23. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, et al. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA 102: 7882–7887.
24. Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, et al. (2007) Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science 316: 240–243.
25. Wiuf C (2006) Consistency of estimators of population scaled parameters using composite likelihood. J Math Biol 53: 821–841.

26. Zhu L, Bustamante CD (2005) A composite-likelihood approach for detecting directional selection from DNA sequence data. Genetics 170: 1411–1421.
27. Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, et al. (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14: 1821–1831.
28. Fischer RA (1922) On the dominance ratio. Proc Roy Soc Edin 55: 399–433.
29. Kimura M (1964) Diffusion models in population genetics. J Appl Probab 1: 177–232.
30. Ewens WJ (2000) Mathematical Population Genetics: I. Theoretical Introduction Springer, 2nd edition.
31. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256–276.
32. Nagylaki T (1980) The strong-migration limit in geographically structured populations. J Math Biol 9: 101–114.
33. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496–1502.
34. Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168: 2373–2382.
35. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
36. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical Recipes: The Art of Scientific Computing Cambridge University Press, 3rd edition.
37. Chang JS, Cooper G (1970) A practical difference scheme for Fokker-Planck equations. J Comput Phys 6: 1–16.
38. Oliphant TE (2006) Guide to NumPy Trelgol Publishing.
39. Oliphant TE (2007) Python for scientific computing. Comput Sci Eng 9: 10–20.
40. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9: 90–95.
41. NIEHS Environmental Genome Project. URL http://egp.gs.washington.edu. Accessed November 6, 2007.
42. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, et al. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2: e286. doi: 10.1371/journal.pbio.0020286.
43. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.
44. Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA 101: 13994–14001.
45. Kumar S, Filipski A, Swarna V, Walker A, Hedges SB (2005) Placing confidence limits on the molecular age of the human-chimpanzee divergence. Proc Natl Acad Sci USA 102: 18842–18847.
46. Kondrashov AS (2002) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21: 12–27.
47. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128: 415–423.
48. Tremblay M, Vézina H (2000) New estimates of intergenerational time intervals for the calculation of age and origin of mutations. Am J Hum Genet 66: 651–658.
49. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083. doi: 10.1371/journal.pgen.1000083.
50. Heinrich JG (2001) Can the likelihood-function value be used to measure goodness of fit? Technical Report 5639, Collider Detector at Fermilab.
51. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, et al. (2007) A genomewide admixture map for Latino populations. Am J Hum Genet 80: 1024–1036.
52. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
53. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, et al. (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74: 979–1000.
54. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR (2009) Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci USA 106: 3871–3876.
55. Weale ME, Weiss DA, Jager RF, Bradman N, Thomas MG (2002) Y chromosome evidence for Anglo-Saxon mass migration. Mol Biol Evol 19: 1008–1021.
56. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.
57. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.
58. Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, et al. (2008) A novel DNA sequence database for analyzing human demographic history. Genome Res 18: 1354–1361.
59. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783–796.
60. Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? Theor Popul Biol 73: 342–348.
61. Pierce DA, Schafer DW (1986) Residuals in generalized linear models. J Am Stat Assoc 81: 977–986.